Data and Metadata. 2024; 3:.548
doi: 10.56294/dm2024.548

ORIGINAL

Detecting hemorrhagic stroke from computed tomographic scans using machine learning models comparison

Detección de ictus hemorrágico a partir de tomografías computarizadas mediante comparación de modelos de aprendizaje automático

Zaynab Boujelb[1,2], Ahmed Idrissi[2,3], Achraf Benba[4], El Mahjoub Chakir[1]

[1] Faculty of Sciences, Ibn Tofail University, Kenitra, Morocco.
[2] Higher Institute of Nursing Professions and Health Techniques (ISPITS), Rabat, Morocco.
[3] Faculty of Sciences, Mohammed V University, Rabat, Morocco.
[4] ENSAM, E2SN, Mohammed V University, Rabat, Morocco.

Cite as: Boujelb Z, Idrissi A, Benba A, Chakir EM. Detecting hemorrhagic stroke from computed tomographic scans using machine learning models comparison. Data and Metadata. 2024; 3:.548. https://doi.org/10.56294/dm2024.548

Submitted: 20-05-2024 **Revised:** 25-08-2024 **Accepted:** 20-12-2024 **Published:** 21-12-2024

Editor: Adrián Alejandro Vitón-Castillo

Corresponding author: Zaynab Boujelb

ABSTRACT

Introduction: stroke is the most leading cause of death and disability worldwide, with hemorrhagic stroke being the most dangerous due to bleeding in the brain. To minimize the impacts, early detection is crucial for effective management and timely intervention. This is precisely the motivation behind our research, which aims to develop a reliable and rapid diagnostic support system.
Method: in this study, the authors combined machine learning (ML) models to detect stroke using a dataset of computerized tomography (CT) images. The study was conducted on a real database containing CT images collected from Moroccan patients. The method used in data organization and preprocessing were performed, followed by feature extraction from each image, such as intensity, grayscale, and histogram characteristics. These extracted features were then compressed using several algorithms, including Principal Component Analysis (PCA). The processed data were fed into the most robust machine learning classifiers based on existing literature.
Results: as a result, the XGBoost model achieved the highest classification accuracy, with 93 % precision, using a Leave-One-Subject-Out (LOSO) validation scheme.
Conclusion: this study is a step forward in improving patient healthcare by enabling early detection, which could lead to timely, potentially life-saving interventions.

Keywords: Hemorrhagic Strokes; Machine Learning; Healthcare; Computed Tomography.

RESUMEN

Introducción: el ictus es la principal causa de muerte y discapacidad en todo el mundo, siendo el ictus hemorrágico el más peligroso debido a la hemorragia cerebral. Para minimizar sus efectos, la detección precoz es crucial para un tratamiento eficaz y una intervención oportuna. Esta es precisamente la motivación de nuestra investigación, cuyo objetivo es desarrollar un sistema de ayuda al diagnóstico fiable y rápido.
Método: en este estudio, los autores combinaron modelos de aprendizaje automático (ML) para detectar accidentes cerebrovasculares utilizando un conjunto de datos de imágenes de tomografía computarizada (TC). El estudio se realizó sobre una base de datos real que contenía imágenes de TC obtenidas de pacientes marroquíes. Se utilizó un método de organización y preprocesamiento de datos, seguido de la extracción de características de cada imagen, como la intensidad, la escala de grises y las características del histograma. A continuación, estas características extraídas se comprimieron mediante varios algoritmos, incluido el análisis

de componentes principales (ACP). Los datos procesados se introdujeron en los clasificadores de aprendizaje automático más robustos basados en la bibliografía existente.
Resultados: como resultado, el modelo XGBoost logró la mayor precisión de clasificación, con un 93 % de precisión, utilizando un esquema de validación Leave-One-Subject-Out (LOSO).
Conclusiones: este estudio es un paso adelante en la mejora de la atención sanitaria de los pacientes al permitir la detección temprana, lo que podría conducir a intervenciones oportunas y potencialmente salvadoras de vidas.

Palabras clave: Accidentes Cerebrovasculares; Aprendizaje Automático; Asistencia Sanitaria; Tomografía Computarizada.

INTRODUCTION

Strokes represent a critical medical emergency, being one of the leading causes of mortality and disability worldwide. Early and accurate diagnosis of strokes is crucial for improving clinical results and reducing sequelae. However, traditional diagnostic methods, which primarily rely on clinical analysis and medical imaging, can be limited by subjectivity and variability in human assessments. For the identification, description, and prognosis of acute strokes, including ischemic and hemorrhagic subtypes, neuroimaging is a crucial technique.[1,2]

Stroke, also known as a cerebrovascular accident (CVA), is a serious medical illness in which the blood flow to a portion of the brain is disrupted or diminished, depriving brain tissue of oxygen and vital nutrients, resulting in cell death. Strokes are primarily classified into two types: ischemic stroke, which accounts for approximately 85 % of cases and occurs when a blood clot obstructs an artery in the brain, and hemorrhagic stroke, which occurs when a blood vessel ruptures, resulting in bleeding in or around the brain. According to[6] intracerebral and subarachnoid hemorrhages are the two main types of hemorrhagic stroke; they are both acute arterial bleeding within the cranial cavity, but they differ in their underlying physiopathology. Depending on which part of the brain is affected, the illness can cause a variety of neurological abnormalities such as movement dysfunction, cognitive problems, and speech issues.[2,3,4,5] Over 16 million people worldwide suffer from stroke each year, making it one of the main causes of death and disability. Stroke also has a substantial financial impact on society and the healthcare system.[7] To mitigate the long-term effects of stroke, early detection and prompt medical care are crucial. This underscores the significance of sophisticated diagnostic technologies, such machine learning, in improving stroke outcomes.[8]

Stroke tops the list of diseases as the second most deadly diseases. In America, about 800 000 of citizens suffer a stroke each year, and over 5,5 million deaths annually.[9,10] For sure, half of stroke patients leaving permanently handicapped and disabled.[10] This disease has an impact economically and it costs over $34 billion yearly.[11] Moving to Maghreb countries, Morocco such an example for beating stroke disease, an epidemiological study revealed that 0,284 % of patients suffering from a stroke disease, with ischemic stroke (IS) constituting 70,9 % of all instances of stroke.[12] In a recent review of stroke, the most etiologies profiles were atherosclerosis and cardioembolic disease, the number of deaths that occur in Moroccans varied in the acute phase from 3 % to 13 %, and the three-month mortality ranged from 4,3 % to 32,5 %. Otherwise, this study estimated the average annual direct cost of managing ischemic stroke to be $3674,32 per patient, with hospitalization costs being the largest component at $1415,06.[13] However, during the last decade, impressive advancements and progress in imaging techniques have been accomplished in symptomatic strategies and therapeutic interventions aimed at mitigating the intense and impact of acute IS, especially with the improvement of revascularization techniques.[14]

Despite technological advancements in medical imaging, diagnosing Cerebrovascular accidents (CVAs) remains a major challenge. Clinicians often need to make rapid decisions based on complex and sometimes ambiguous images, which can lead to diagnostic errors or delays in treatment. Therefore, it is imperative to develop more precise and automated tools to assist physicians in this critical task.[15] With the emergence of artificial intelligence (AI), new opportunities arise to alter medical diagnosis with accuracy, realistically, the AI methods and applications for imaging acute (CVAs) disease have been performed, involving tools for classification, quantification, monitoring and prediction.

The objective of this work is to enhance the detection of hemorrhagic stroke and assist in diagnosis and reliability for clinicians. Hemorrhagic stroke is a critical condition that requires rapid and accurate diagnosis to improve patient outcomes. Current diagnostic methods can be time-consuming and prone to errors, which can delay treatment. This study aims to address these issues by providing fast and accurate analysis and interpretation of brain images using the CT scanner modality. Several machine learning algorithms with

different parameters were used, and a comparison was conducted to identify the most suitable one for this application. XGBoost demonstrated the highest accuracy, ensuring its applicability in neurology.

METHOD

The engagement of stroke prediction analysis is very important in the medical field because the timely and correct assessment of individuals at risk of stroke can greatly reduce the adverse effects of stroke. The use of machine learning has significantly improved the accuracy and reliability of predictive models. In this regard, this study aims to differentiate between healthy and stroke subjects from CT images using various image processing, dimensionality reduction and machine learning techniques. Each step, from organizing the data to analyzing the results, is crucial to achieving our goal. In the process of developing models, data becomes ready for us when it has been processed. There is a need for a preprocessed dataset and techniques of machine learning for carrying out the model construction. In addition, this study is a retrospective observational study aimed at comparing the performance of various machine learning models in detecting hemorrhagic stroke from computed tomographic (CT) scans. The study was conducted in accordance with the ethical standards of the institutional and national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Consent was waived due to the retrospective nature of the study. All patient data were anonymized to ensure confidentiality.

The methodology followed in our study is schematized by a block diagram presented below (figure 1) with details.

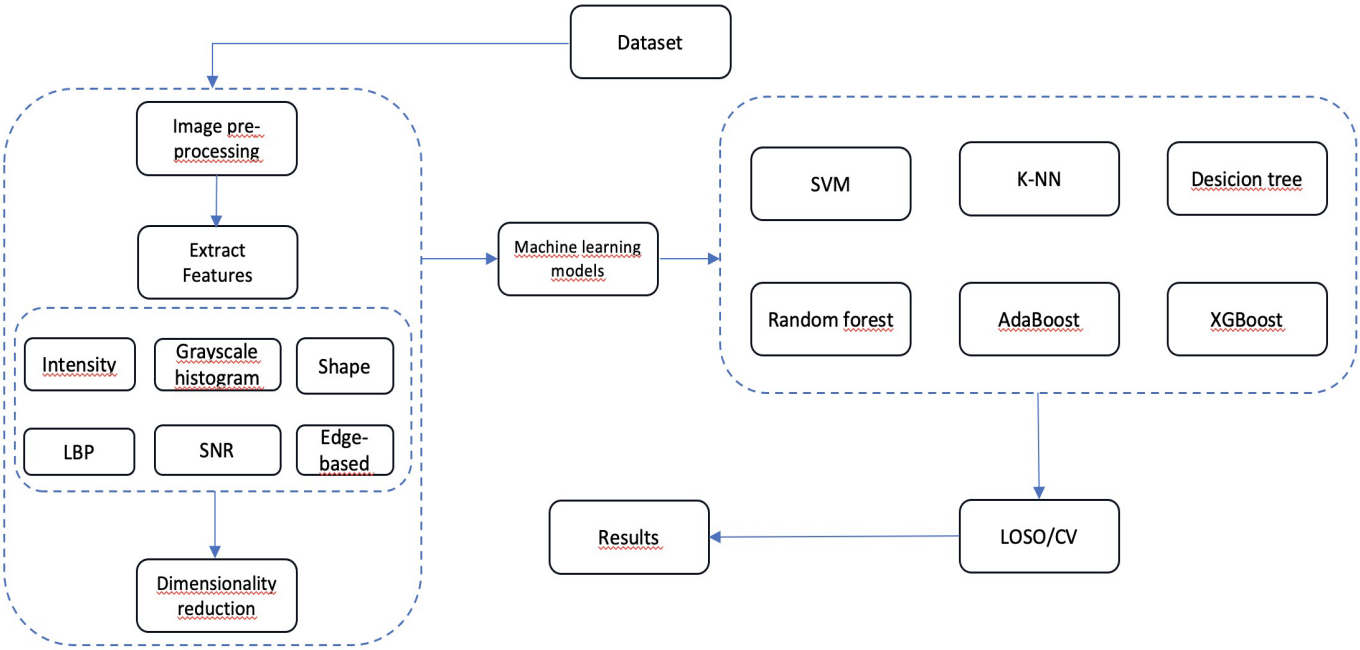

Figure 1. Block diagram

Dataset

The stroke prediction dataset, was conducted at a Moroccan public hospital located in Rabat. The data collection and analysis were carried out between May 2024 and September 2024, which was gathered over the course of four months from 85 Moroccan patients, was used in the study. 266 distinct CT scans from new patients and the Picture Archiving and Communication System (PACS) are included in this data. In order to make processing and analysis easier, the initial step was to arrange all of the CT scans in a systematic manner. CT scans from figure 2 were divided into two groups: “NORMAL,” which included 22 healthy participants, and “SICK,” which is referred to as “MALADE” in figure 2 and included 63 patients with CVA diagnoses. In order to establish distinct labels for the classification model, each group consists of a collection of photos reflecting extremely different instances. These images are separated into 64 normal CT scans and sick patients, with 0 denoting “NORMAL” and 1 denoting “SICK.”

The second step was folder structure, each main folder is made up of subfolders, with each subfolder representing a single subject. Each sub-folder can contain one or more CT images of the same subject (figure 3). The diversity of images per subject provides a more complete representation of each patient’s condition. This subject-based organization is important to ensure consistent data analysis and to correctly apply cross-validation (avoiding mixing images from the same subject between training and testing).

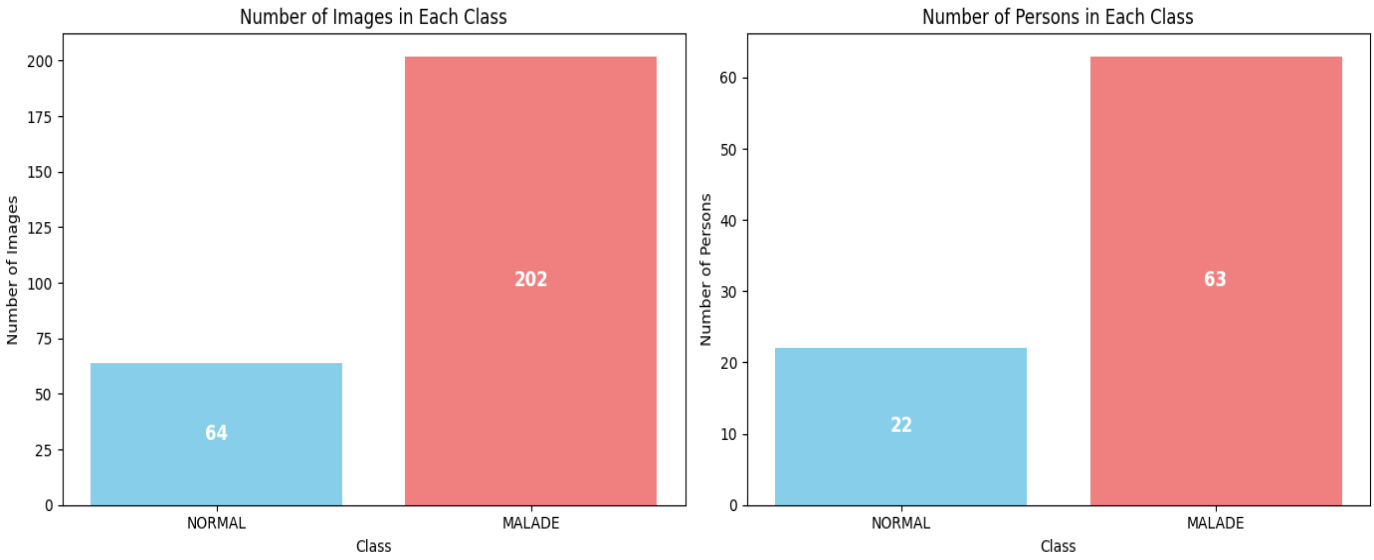

Figure 2. Dataset organization

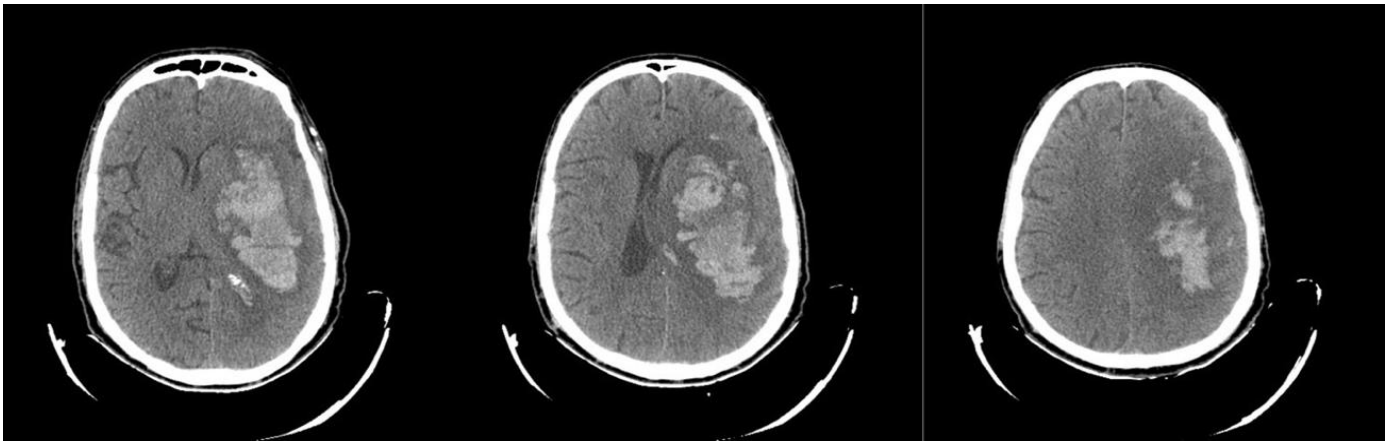

Figure 3. An example of sub-folder from database used, same patient with hemorrhagic stroke CT images (axial section)

Pre-processing

In this work used for stroke detection, pre-processing played a crucial role in preparing the data for machine learning models. It represents an important technique for improving the precision of the machine learning models by reducing noise, ensuring proper feature scaling, and optimizing the execution of algorithms.[18,19,20,21] Several preprocessing steps are implemented to prepare the data for modeling. All images were resized to a uniform size of 128x128 pixels. This standardizes the dimensionality of the images, making feature extraction simpler and avoiding problems associated with size variance. In addition, images were converted to grayscale to focus only on intensity information, which is relevant for brain tissue analysis. It also reduces the complexity of data without losing important information.

Feature extraction

After pre-processing, a lot of information is still contained in the image or signal data. Feature extraction concentrate on choosing the best parts of the image that will enable the model to learn. Furthermore, it is a key step in transforming images into digital vectors that can be processed by classification models.[22,23] These could include edges, textures, shapes, or colors in an image. For each image, various features have been extracted to represent several aspects of the CT image. In our study, the authors focused on intensity characteristic, grayscale histogram, Local Binary Pattern (LBP), shape characteristics, signal to noise ratio (SNR) and edge-based features.

Intensity characteristic

One of the image fundamental properties is intensity, which is translated from pixels. In this context, it refers to how bright or dark a given point is. It is also an important characteristic in many image processing and computer vision problems. Algorithms can segment objects, detect objects, enhance the resolution of

images, analyze textures, and perform feature extraction suitable for machine learning applications by looking at the changes in intensity. The common intensity – based features are, mean, variance, standard deviation, histogram, moments and entropy. In addition, these features allow for the quantitative depiction of the image internal visual structure, which is very helpful in operations like edge detection, object recognition, and image classification.[24] Moreover, there is two main characteristics in our study which is the average intensity and the intensity standard deviation. In a CT scan, the average of all the pixels is reflective of the general luminosity of the image. This characteristic tells whether the image is on the lighter or darker side in general, which could be a measure of the variation in tissue density (e.g., dead tissue vs. normal tissue). Whereas, the standard deviation determines how spread out the intensities of the pixels are with respect to the mean. A high standard deviation values results reveals a high variability in intensity, usually caused by irregularities like lesions. This property assists greatly in detecting any abnormalities in tissues.

Grayscale histogram

A grayscale histogram is used in image analysis, which is a simplified profile of the distribution of pixel intensity that suggests information about the brightness, contrast, and texture of the image. This is why it is often utilized in many applications such as object detection, image splitting and sorting. A histogram almost for all intents and purposes points out the number of pixels that each possible intensity level has in the image, and in the process gives a precise overview of the intensity values contained in the image.[25] Whitin the scope of this study, the grayscale histogram is constructed for each image with 16 equal sized bins per image contributing to a gray range (0-255) division into 16 parts. The histogram describes distribution of the gray levels thus enabling texture description in general. For instance, given an image with a large block of high gray levels, one may imply that the structure is quite dense.

Local Binary Pattern (LBP)

LBP (Local Binary Patterns) is a method that is used to analyze and extract local patterns from an image. It refers to a commonly used metric for determining the texture of an image in the field of computer vision. The LBP technique has found quite a number of applications in texture classification, object detection and face recognition especially for its insensitivity to changes in light and ability to capture textures of the face.[26,27] LBP achieves this in this work by presenting the statistical information of the spatial structure of an image. For every pixel present in the image, LBP takes into account the neighboring values and creates a binary code for the corresponding defined pattern. As a result, it's possible to numerically describe given textures in relations to repetitive and specific patterns for instance micro-Architecture of normal brain which is very important in identification of such pathology as micro-injuries.

Shape characteristics

The characteristics of shapes are a category of tools in image processing and pattern recognition that helps to define the shape of a certain object or region in an image. Such features include important details regarding the edges, outlines, and shape of the object in question. Such shape descriptors are extremely useful in applications including object recognition, classification, or segmentation, as they make it possible to differentiate between various shapes or structures contained in an image. Form features have a broad usage in medical imaging, in which they are also used to define the shape of tumors and other biological structures, and in image retrieval systems, which index objects according to their contours.[26] The two properties of this shape are area and perimeter, the area represents the number of pixels that belong to a detect region, such as a brain lesion. A large area may indicate a large damaged cerebral region, while a small one indicates a more localized zone. The second one is perimeter which is measure the length of a contour of the defined area and this provides clearer picture of the structure and its elaborateness. A cerebral region with much irregularities will have a perimeter that is larger than average which is indicative of certain pathologies.

Signal-to-noise ratio (SNR)

In the context of image processing, SNR is important due to the fact that excess noise can cover important aspects of the image, making information extraction very hard. SNR is also enhanced by the use of pre-processing techniques such as filtering or using noise reduction algorithms. For instance, in medical imaging, one would aim to produce a high SNR for a clear depiction of internal body parts which allows for making more confident and easier decision. On the other hand, a low SNR means getting distorted images or worse signal quality, which can adversely affect detection and analysis, it also causes doubts and more chances of perceptional misinterpretation.[29] With high-SNR data, training machine learning models becomes easier as the trained features are more representative and therefore, generalize better across different tasks improving overall performance.

Furthermore, the signal-to-noise ratio is measured as the ratio of the average of the image to its standard deviation, in this study, SNR measures how well the useful information is present in CT scans of brain as

compared to the noise and it is directly affecting the accuracy, reliability, and clarity of the results.

Edge-based features

Edge-based features refer to a section of image features that seek to understand the information within the images in terms of their boundaries and the edges present in that particular image. The changes in contrast that form edges are of great concern and are usually linked to object contouring, surface designs or texture variations among other things. Therefore, edge detection becomes one of the fundamental processes of image processing due to the fact that edges carry important information on the outline and shape of the objects present in the image.[27] In order to enhance the overall image quality of brain CT scans, a Sobel filter has been implemented to Extract Edges. The total values derived exemplifies the degree of sharpness or details in an image. For instance, an image that contains many intricate structures will yield a high sum, which realistically offers a great difference when comparing normal tissue (less details) with pathological tissues (more details).

Dimensionality reduction

Dimensionality is defined as the feature or variable count in a dataset. In most cases, when extracting the relevant features from the data, it is also necessary to reduce the dimensions of the dataset while preserving the essential content. Considering data with high-dimensional characteristics, it becomes much trivial to start the analysis because of having the "curse of dimensionality", which simply results in increased costs of computation and reducing returns in the performance level of the model. Dimensionality reduction techniques in this respect are those that seek to reduce the dimensions of the data to the least number without losing a lot of information in the process. This is very important since it helps in enhancing the performance of the model as well as aiding the comprehension and representation of the data. Algorithms like principal component analysis (PCA) help to achieve dimensionality reduction by finding the prominent features of the data.[30,31,32] In adding, CT scan-derived brain image features have a lot of dimensions which can make it hard to train a model without falling into overfitting issues.

One of the popular techniques for reducing the dimensionality of the data is Principal Component Analysis (PCA). Principal components are linear combinations of the original variables, which are uncorrelated with each other, and are arranged in the decreasing order of variance explained by them. PCA can be evaluated by implementing algorithms such as eigen decomposition and SVD. Importantly, the principal components are also orthogonal which makes it even simpler in the interpretation and visualization of the results.[33,34,35,36,37,38]

These components maximize the variance of the data while reducing the number of dimensions. In this study, PCA was applied to retain 95 % of the total variance, thus simplifying the data while retaining its relevance for classification.

Kernel Principal Component Analysis (KPCA) can be treated as an advanced version of traditional PCA techniques as KPCA performs analysis on non-linear data structures of the given set. KPCA is extensively utilized for reducing dimensions, thus facilitating the understanding and handling of high dimensional datasets. It works well with data containing nonlinearities, such as images and complex scientific data.[38]

In this case, it employs kernel functions such as RBF which is a Gaussian, to map the data into a higher dimensional space where separation is easier. This is helpful especially in cases where features are not linearly related. Such as often observed in CT images, where complex structures are present.

Classification

Once dimensionality has been reduced, several classification algorithms have been used to differentiate between healthy and stroke subjects, like SVM, k-NN, Decision Tree, Random Forest, AdaBoost and XGBoost.

SVM (Support vector machines) is a supervised machine-learning algorithm for various classification and regression problems, relying on statistical learning theory, as well as on ideas from convex optimization.[39]

To identify suitable class boundaries, SVMs with various kernels: (linear, polynomial, RBF, sigmoid) were applied. The selection of kernel allows relationships between features to be either turned into linear or non-linear as the case demands. For example, the RBF kernel is typically applied to data with complex distributions.

For, K-Nearest Neighbors (KNN) is a type of instance-based learning, where KNN takes a test sample and assigns its classification based on the classifications of training examples that are 'close' to the given test sample using a distance measure. This distance measure is typically the Euclidean distance, which is then used to find the k nearest training samples of the test instance and the class which is most common amongst those k nearest neighbor training samples is assigned to the test instance.[40,41,42]

In this case, each observation for k-NN classification is classified based on its k-n neighbors, where the neighbors are defined in terms of distance from the observation. Different values of k (1, 3, 5, 7, 9) were examined to ascertain the performance of the model. It is a straightforward method; however, it can be affected by the number of dimensions, hence the importance of prior reduction before analysis.

A decision tree can be represented in a way similar to that of a tree in which each non-leaf node corresponds

to a conditional test concerning some attribute, every edge corresponds to the outcome of the test performed at the parent node and every terminal node is associated with a prediction or a class label.[43,44]

Moreover, various decision trees with the depths (3, 5, 10) were built. Recalling that in every node of the tree there is a feature question used to divide the data into two classes (our study). The higher the depth of the tree, the better its ability to model more complex relationships, but this may also be a cause of overlearning.

Passing to Random Forest which is a supervised learning algorithm, at the core, that evaluates a large number of decision trees created at training time – and gives prediction based on the class that is most frequent among them (classification) or on the average prediction made by all such trees (regression).[45,46,47]

In essence, a Random Forest comprises multiple decision trees, which cast votes for the most probable classification. In our model, there have been implemented various numbers of trees (50, 100, 200). This approach is not prone to overlearning due to the number of trees that "democratize" for the result.

Regarding the two models, AdaBoost (Adaptative Boost) and XGBoost (Xtreme Gradient Boost), there are many differences between the two models on several properties such as: the mechanism, the strengths and limitations of each algorithm.

For AdaBoost, it combines numerous weak classifiers to build a strong classifier by iteratively modifying the weights of misclassified instances, prioritizing tough situations in following rounds. It is especially successful at minimizing bias and variance, making it useful for enhancing the performance of various base classifiers, such as neural networks and fuzzy inference systems. AdaBoost can be computationally demanding and may struggle with noisy data, but improvements such as GPU acceleration and noise-resistant versions have been developed to solve these concerns.[48,49]

XGBoost is a prediction algorithm aimed at improving accuracy by building trees in a sequential manner one after another whereby each new tree corrects the consequences and faults of the precedent one. It so adds upon the performance by using a gradient boosting framework. XGBoost is famous for being very accurate and efficient it can manage large-scale data sets with complex models without any difficulties. It incorporates regularization techniques in order to tame this problem and is tolerant to various forms and types of data and their distributions.[50,51,52] Usage of XGBoost can however be effective boosting algorithm for classification tasks.

In this work, AdaBoost model is iteratively improved to correct flaws made by previous estimators. Performance was analyzed with 50 and 100 estimates. Conversely to XGBoost, which is evaluated with learning rates of 0,01, 0,1, and 0,2. The parameter known as the learning rate determines how quickly a model modifies its parameters. It is said that XGBoost can perform well on high-dimensional datasets and is robust to noise and outliers.

RESULTS

In this study, kernel Principal Component Analysis (KPCA) was used as a feature selection method to compare different machine learning models, which were then evaluated using Leave-One-Subject-Out Cross-Validation (LOSO-CV) (figure 4). Intricate, non-linear correlations were extracted from the data using KPCA, providing a more sophisticated feature transformation than conventional PCA. The assessment guarantees that model performance is evaluated in a subject-independent manner by integrating LOSO-CV, offering insights on generalization across other subjects. This strategy is especially pertinent to domains where subject variability may affect model robustness, such as neurology and healthcare. The comparative analysis demonstrated that distinct models display differing degrees of susceptibility to feature transformation and cross-validation techniques, emphasizing the significance of choosing a model based on the task at hand.

The models were evaluated using the Leave-One-Subject-Out (LOSO) cross-validation method, which involves training on all subjects except one, and testing on the excluded subject. This approach uses one patient's data (CT scans of hemorrhagic stroke) as a test set, while the other patients' data are used for training, ensuring optimized results. This method ensures that each patient is adaptively used as a test set. Among all the tested models shown in table 2 and figure 5, XGBoost performed best with a learning rate of 0,2, achieving an accuracy of 93 %. This means that the model correctly classified 93 % of the subjects as normal or diseased. The remarkable sensitivity of 98 % means that 62 patients were correctly identified and only one patient was missed, demonstrating high accuracy in detecting stroke. The specificity of the model was 77 %, with 17 normal subjects correctly diagnosed. This high accuracy demonstrates the efficiency of XGBoost in managing complex data and capturing the relationships between features extracted from scanner images. Challenges in the medical field include the risk of misclassifying a diseased subject as normal, which can be dangerous. Another algorithm, k-NN (k=9), also showed a sensitivity of 98 %, but only 11 normal subjects were correctly diagnosed. The use of dimensionality reduction via PCA and Kernel PCA contributed to improved model performance by limiting noise and data complexity. XGBoost is widely recognized for its accuracy, speed, and versatility in various fields, including imaging in neurology, and it consistently ranks as one of the top-performing algorithms in machine learning.

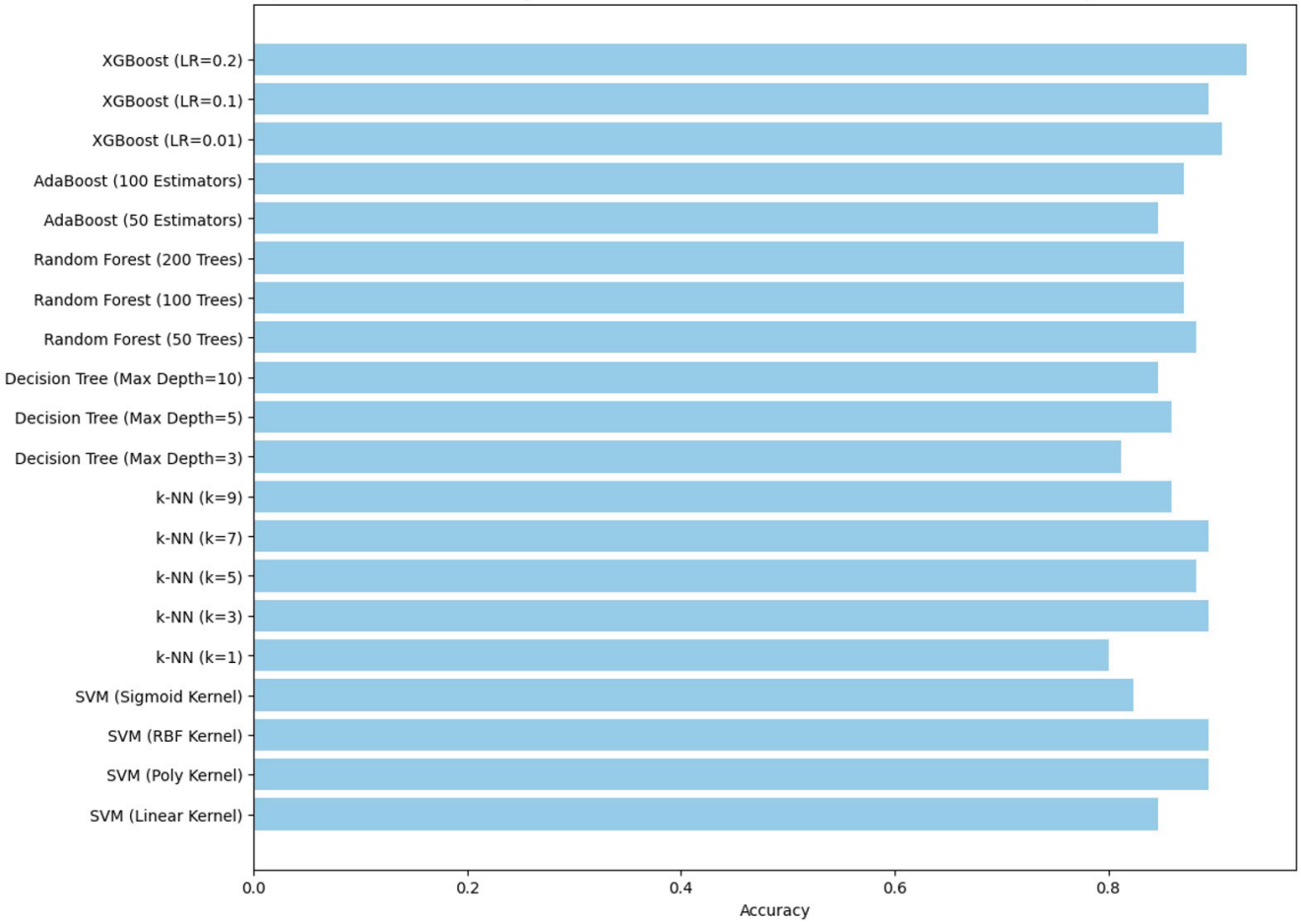

Figure 4. Comparison of machine learning models with kernel PCA selected features (LOSO/Cross Validation)

Table 1. Metrics of tested models

Classifier	Accuracy	Sensitivity (Recall)	Specificity	TP	TN	FP	FN
SVM (Linear Kernel)	0,85	0,84	0,86	53	19	3	10
SVM (Poly Kernel)	0,89	0,95	0,73	60	16	6	3
SVM (RBF Kernel)	0,89	0,92	0,82	58	18	4	5
SVM (Sigmoid Kernel)	0,82	0,84	0,77	53	17	5	10
k-NN (k=1)	0,8	0,89	0,55	56	12	10	7
k-NN (k=3)	0,89	0,97	0,68	61	15	7	2
k-NN (k=5)	0,88	0,95	0,68	60	15	7	3
k-NN (k=7)	0,89	0,95	0,73	60	16	6	3
k-NN (k=9)	0,86	0,98	0,5	62	11	11	1
Decision Tree (Max Depth=3)	0,81	0,86	0,68	54	15	7	9
Decision Tree (Max Depth=5)	0,86	0,92	0,68	58	15	7	5
Decision Tree (Max Depth=10)	0,85	0,92	0,64	58	14	8	5
Random Forest (50 Trees)	0,88	0,95	0,68	60	15	7	3
Random Forest (100 Trees)	0,87	0,95	0,64	60	14	8	3
Random Forest (200 Trees)	0,87	0,97	0,59	61	13	9	2
AdaBoost (50 Estimators)	0,85	0,9	0,68	57	15	7	6
AdaBoost (100 Estimators)	0,87	0,94	0,68	59	15	7	4
XGBoost (LR=0,01)	0,91	1	0,64	63	14	8	0
XGBoost (LR=0,1)	0,89	0,97	0,68	61	15	7	2
XGBoost (LR=0,2)	0,93	0,98	0,77	62	17	5	1

https://doi.org/10.56294/dm2024.548

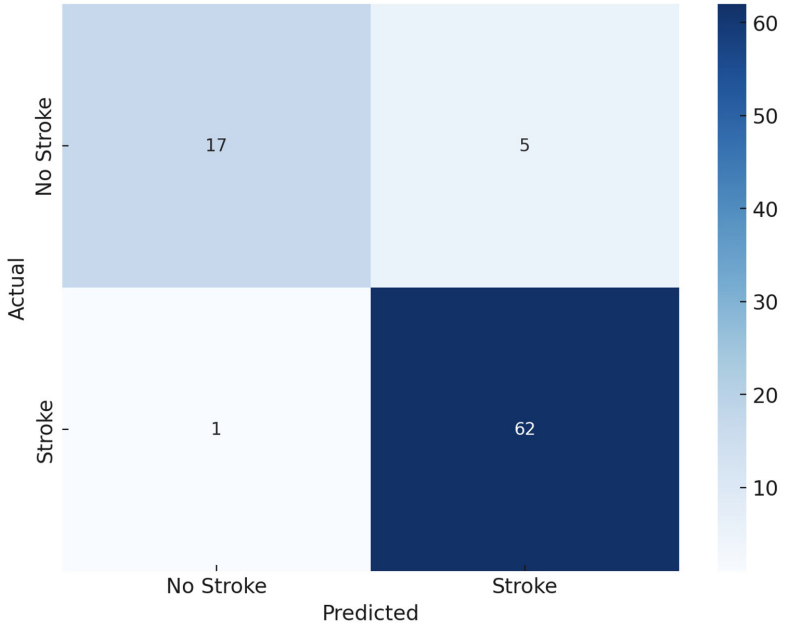

Figure 5. Confusion Matrix for XGBoost (LR=0,2)

DISCUSSION

The recent progression of AI and Machine Learning has revolutionized stroke imaging, detecting, segmenting, and classifying different types of strokes, more so in the case of ischemic and hemorrhagic strokes. By examining enormous datasets to find trends and enhance patient results, machine learning (ML) is transforming the diagnosis and prognosis of stroke. In their study of machine learning applications for stroke prediction, authors [16] emphasized methods like Random Forest (RF) and Support Vector Machines (SVM), which were shown to be the best in a number of studies. The evaluation divided 39 papers published between 2007 and 2019 into four categories: prognostication, diagnosis, treatment, and stroke prevention. SVM models produced the best accuracy, especially in assignments involving the diagnosis of strokes, where MRI and CT imaging were regularly used. In [7] and using clinical data from Bangladeshi hospitals, researchers used a weighted voting classifier that incorporated various machine learning models and achieved a 97 % accuracy in stroke prediction. Similar to this, in [8] researchers used a variety of algorithms, including Random Forest, Decision Trees, and Logistic Regression; Random Forest achieved an accuracy rate of 96 %. All of these researches show how machine learning (ML) can be used to identify strokes early on, choose the best course of treatment, and enhance long-term patient outcomes. Advanced imaging methods and ensemble models yield better predictive performance.

The study conducted by researchers in [17] leveraged a database, which includes three datasets with 212 stroke instances and a range of non-stroke examples (52, 69, 79). They employed the Naïve Bayes used for handling datasets with multiple attributes, J48 decision tree for classification based on rules derived from data, k-Nearest neighbors (K-NN) for classification based on feature similarity and Random Forest which is based on decision trees. The performance of the models was evaluated using metrics. The Random Forest and J48 algorithms achieved the highest accuracy of 99,8 %, while Naive Bayes achieved an accuracy of 85,6 %. The models demonstrated strong performance, particularly for classifying ischemic and hemorrhagic strokes, with k-NN and Random Forest models providing precise and reliable predictions. Otherwise, the authors in [8] have concluded that the Random Forest algorithm achieved the best performance in predicting strokes, with an accuracy of 96 %, followed by the Decision Tree classifier with 94 % accuracy, the Voting Classifier with 91 %, and Logistic Regression with 87 %. The authors used SMOTE (Synthetic Minority Over-sampling Technique) to handle the class imbalance in the dataset, which was necessary since only 249 rows indicated stroke, while 4861 rows did not. The Random Forest algorithm emerged as the most robust model, other models such as AdaBoost, SVM, or XGBoost models can be more precisive by using larger datasets. In [7] XGBoost model achieved a high prediction accuracy of 96 % which performed by dataset that contains 5,110 records of patients.

The current study showed that an XGBoost model achieved a classification accuracy of 93 %, sensitivity of 98 %, and specificity of 77 % in hemorrhagic stroke detection from CT scans. This is comparable to other studies like that of researchers using weighted voting classifier achieving a 97 % accuracy in stroke prediction using clinical data from Bangladesh hospitals. In another study Random Forest, an efficacy of 96 % is reported. The use of KPCA in our feature selection and LOSO-CV model evaluation here ensures strong and generalizable results. In contrast, the study by researchers using Naïve Bayes, J48, k-NN, and Random Forest achieved the highest accuracy of 99,8 % with Random Forest and J48. This shows the effect of advanced machine learning

techniques and feature selection methods in improving stroke diagnosis, with our study contribution on how XGBoost can be used in such an exercise.

CONCLUSIONS

In summary, this study confirmed the effectiveness of machine learning algorithms, especially the XGBoost classifier, in improving the early diagnosis of hemorrhagic stroke with the use of CT scans. Through efficient construction of the dataset, preprocessing, and data dimensionality reduction, a satisfactory classification accuracy of 93 % was achieved, confirming the strength of the technique. The Leave-One-Subject-Out (LOSO) validation scheme used in the model further increases the reliability of the model. This method aims to aid stroke management by reducing time and increasing diagnostic accuracy, a key aspect that directly impacts patients' clinical outcomes. Future perspectives may include dataset expansion and model improvement to achieve even higher accuracy and relevance in different clinical situations.

REFERENCES

1. Benjamin EJ, Muntner P, Alonso A, Bittencourt MS, Callaway CW, Carson AP, et al. Heart Disease and Stroke Statistics—2019 Update: A Report From the American Heart Association. Circulation [Internet]. 2019 Mar 5;139(10). https://doi.org/10.1161/cir.0000000000000659

2. Feigin VL, Forouzanfar MH, Krishnamurthi R, Mensah GA, Connor M, Bennett DA, et al. Global and regional burden of stroke during 1990–2010: findings from the Global Burden of Disease Study 2010. The Lancet. 2014 Jan;383(9913):245–55. https://doi.org/10.1016/s0140-6736(13)61953-4

3. Benjamin EJ, Blaha MJ, Chiuve SE, Cushman M, Das SR, Deo R, et al. Heart Disease and Stroke Statistics—2017 Update: A Report From the American Heart Association. Circulation [Internet]. 2017 Mar 7;135(10). https://doi.org/10.1161/cir.0000000000000485

4. Campbell BCV, De Silva DA, Macleod MR, Coutts SB, Schwamm LH, Davis SM, et al. Ischaemic stroke. Nature Reviews Disease Primers. 2019 Oct 10;5(1). https://doi.org/10.1038/s41572-019-0118-8

5. Donnan GA, Fisher M, Macleod M, Davis SM. Stroke. Lancet [Internet]. 2008;371(9624):1612–23. https://doi.org/10.1016/S0140-6736(08)60694-7

6. Morais Filho AB de, Rego TL de H, Mendonça L de L, Almeida SS de, Nóbrega ML da, Palmieri T de O, et al. The physiopathology of spontaneous hemorrhagic stroke: a systematic review. Reviews in the Neurosciences [Internet]. 2021 Feb 15; https://doi.org/10.1515/revneuro-2020-0131

7. Emon MU, Rahman MM, Ferdousi R, et al. Performance analysis of machine learning approaches in stroke prediction. IEEE Xplore. 2020. Available from: https://ieeexplore.ieee.org/document/9297525

8. Tazin T, Alam MN, Dola NN, Bari MS, Bourouis S, Monirujjaman Khan M. Stroke Disease Detection and Prediction Using Robust Learning Approaches. Journal of Healthcare Engineering [Internet]. 2021 Nov 26;2021:e7633381. https://doi.org/10.1155/2021/7633381

9. Ovbiagele B, Nguyen-Huynh MN. Stroke Epidemiology: Advancing Our Understanding of Disease Mechanism and Therapy. Neurotherapeutics [Internet]. 2011 Jun 21;8(3):319–29. https://doi.org/10.1007/s13311-011-0053-1

10. Donkor ES. Stroke in the 21st century: A snapshot of the burden, epidemiology, and quality of life. Stroke Research and Treatment. 2018 Nov 27;2018(3238165):1–10. https://doi.org/10.1155/2018/3238165

11. Boehme AK, Esenwa C, Elkind MSV. Stroke Risk Factors, Genetics, and Prevention. Circulation research [Internet]. 2017 Feb 3;120(3):472–95.

12. Engels T, Baglione Q, Audibert M, Viallefont A, Mourji F, El Alaoui Faris M. Socioeconomic Status and Stroke Prevalence in Morocco: Results from the Rabat-Casablanca Study. Ikram MA, editor. PLoS ONE [Internet]. 2014 Feb 28;9(2):e89271. https://doi.org/10.1371/journal.pone.0089271

13. Kharbach A, M. Obtel, L. Lahlou, Jehanne Aasfara, Nour Mekaoui, Rachid Razine. Ischemic stroke in Morocco: a systematic review. BMC Neurology. 2019 Dec 30;19(1). https://doi.org/10.1186/s12883-019-1558-1

14. Herpich F, Rincon F. Management of Acute Ischemic Stroke. Critical Care Medicine [Internet]. 2020 Oct 9;48(11):1654–63. https://doi.org/10.1097/ccm.0000000000004597

15. Soun JE, Chow DS, Nagamine M, Takhtawala RS, Filippi CG, Yu W, et al. Artificial Intelligence and Acute Stroke Imaging. American Journal of Neuroradiology [Internet]. 2021 Jan 1;42(1):2–11. https://doi.org/10.3174/ajnr.A6883

16. Sirsat MS, Fermé E, Câmara J. Machine Learning for Brain Stroke: A Review. Journal of Stroke and Cerebrovascular Diseases [Internet]. 2020 Oct 1;29(10):105162. https://doi.org/10.1016/j.jstrokecerebrovasdis.2020.105162

17. Shoily TI, Islam T, Jannat S, Tanna SA, Alif TM, Ema RR. Detection of Stroke Disease using Machine Learning Algorithms. 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT). 2019 Jul; https://doi.org/10.1109/icccnt45670.2019.8944689

18. Kotva Goudoungou S, Dayang P, Tchomte ND, Ngossaha JM, Moffo FM, Mitton N. Covid-19 Data Preprocessing Approach in Machine Learning for Prediction. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. 2024;328–44. https://doi.org/10.1007/978-3-031-56396-6_21

19. Kazi S, Vakharia P, Shah P, Gupta R, Tailor Y, Mantry P, et al. Preprocessy: A Customisable Data Preprocessing Framework with High-Level APIs [Internet]. IEEE Xplore. 2022 [cited 2023 Jun 17]. p. 206–11. https://doi.org/10.1109/CDMA54072.2022.00039

20. Cuevas E, Rodríguez AN. Image Processing and Machine Learning, Volume 1. 2024. https://doi.org/10.1201/9781003287414

21. Cuevas E, Rodríguez AN. Image Processing and Machine Learning, Volume 2. 2024 Jan 2; https://doi.org/10.1201/9781032662466

22. Zarinabegam Mundargi, Bhatti S, Chandra A, Kamble A, Bijin Jiby, Rohit Arole. PrePy - A Customize Library for Data Preprocessing in Python. 2023 Jan 24; https://doi.org/10.1109/iconat57137.2023.10080134

23. Miquilini P, Barros RC, de V, Basgalupp MP. Enhancing discrimination power with genetic feature construction: A grammatical evolution approach. 2022 IEEE Congress on Evolutionary Computation (CEC). 2016 Jul 1; https://doi.org/10.1109/cec.2016.7744274

24. Gara M, Tasi TS, Péter Balázs. Machine Learning as a Preprocessing Phase in Discrete Tomography. Lecture notes in computer science. 2012 Jan 1;109–24. https://doi.org/10.1007/978-3-642-32313-3_8

25. Halder TK, Sarkar K, Mandal A, Sarkar S. A novel histogram feature for brain tumor detection. International Journal of Information Technology. 2022 Apr 4;14(4):1883–92. https://doi.org/10.1007/s41870-022-00917-w

26. Tunuri Sundeep, Uppalapati Divyasree, Karumanchi Tejaswi, Ummadi Reddy Vinithanjali, Anumandla Kiran Kumar. Feature Extraction of Ophthalmic Images Using Deep Learning and Machine Learning Algorithms. Engineering Proceedings [Internet]. 2023 Oct 26 [cited 2024 Aug 31];56(1). https://doi.org/10.3390/asec2023-15231

27. Korichi M, Meraoumia A, Aiadi KE. Improved biometric identification system using a new scheme of 3D local binary pattern. International Journal of Information and Communication Technology. 2019;14(4):439. https://doi.org/10.1504/ijict.2019.101863

28. Rajan AP, Mathew AR. Evaluation and Applying Feature Extraction Techniques for Face Detection and Recognition. Indonesian Journal of Electrical Engineering and Informatics (IJEEI). 2019 Dec 4;7(4). https://doi.org/10.11591/ijeei.v7i4.935

29. Almohamad TA, Mohd Salleh MF, Mahmud MN, Sa'D AHY. Simultaneous Determination of Modulation Types and Signal-to-Noise Ratios Using Feature-Based Approach. IEEE Access. 2018;6:9262–71. https://doi.org/10.1109/access.2018.2809448

30. Sun Z, Xing W, Guo W, Kim S, Li H, Li W, et al. A Survey on Dimension Reduction Algorithms in Big Data Visualization. Springer eBooks. 2020 Jan 1;375–95. https://doi.org/10.1007/978-3-030-48513-9_31

31. Sharma N, Saroha K. A novel dimensionality reduction method for cancer dataset using PCA and Feature Ranking. 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI). 2015 Aug; https://doi.org/10.1109/icacci.2015.7275954

32. Kvinge H, Farnell E, Kirby M, Peterson C. Monitoring the shape of weather, soundscapes, and dynamical systems: a new statistic for dimension-driven data analysis on large datasets. 2021 IEEE International Conference on Big Data (Big Data). 2018 Dec 1;41:1045–51. https://doi.org/10.1109/bigdata.2018.8622365

33. Abdi H, Williams LJ. Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics. 2010 Jun 30;2(4):433–59. https://doi.org/10.1002/wics.101

34. Chowdhury A, Bose A, Zhou S, Woodruff DP, Petros Drineas. A Fast, Provably Accurate Approximation Algorithm for Sparse Principal Component Analysis Reveals Human Genetic Variation Across the World. Lecture notes in computer science. 2022 Jan 1;13278:86–106. https://doi.org/10.1007/978-3-031-04749-7_6

35. Salem N, Hussein S. Data dimensional reduction and principal components analysis. Procedia Computer Science. 2019;163:292–9. https://doi.org/10.1016/j.procs.2019.12.111

36. Hasan MM, Bala B, Atsuo Yoshitaka. SVD aided eigenvector decomposition to compute PCA and it's application in image denoising. 4th International Conference on Informatics, Electronics and Vision, ICIEV 2015. 2015 Jun 1;1–6. https://doi.org/10.1109/iciev.2015.7334007

37. Wang Dongshu. Linear Projection Based Dimension Reduction Analysis in Algebra Space. Jisuanji gongcheng. 2005 Jan 1;

38. Mallegowda M, Tanupriya R, Vishnupriya C, Kanavalli A. Serial vs parallel execution of Principal Component Analysis using Singular Value Decomposition. Proceedings of the 2nd International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics, ICIITCEE 2024. 2024 Jan 24; https://doi.org/10.1109/iitcee59897.2024.10467693

39. Niu Y, Ye S. Data Prediction Based on Support Vector Machine (SVM)—Taking Soil Quality Improvement Test Soil Organic Matter as an Example. IOP Conference Series: Earth and Environmental Science. 2019 Jul 1;295(2):012021. https://doi.org/10.1088/1755-1315/295/2/012021

40. Khumukcham Robindro, Singh YR, Clinton UB, Linthoingambi Takhellambam, Hoque N. CD-KNN: A Modified K-Nearest Neighbor Classifier with Dynamic K Value. Lecture notes in electrical engineering. 2022 Jan 1;753–62. https://doi.org/10.1007/978-981-19-4831-2_62

41. Niu Y, Wang X. On the k-Nearest Neighbor Classifier with Locally Structural Consistency. Lecture notes in electrical engineering. 2013 Oct 11;269–77. https://doi.org/10.1007/978-3-642-40630-0_34

42. Chen S. K-Nearest Neighbor Algorithm Optimization in Text Categorization. IOP Conference Series: Earth and Environmental Science. 2018 Jan;108:052074. https://doi.org/10.1088/1755-1315/108/5/052074

43. Mitu MM, Arefin S, Saurav Z, Hasan MdA, Farid DMd. Pruning-Based Ensemble Tree for Multi-Class Classification. 2024 6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT). 2024 May 2;481–6. https://doi.org/10.1109/iceeict62016.2024.10534584

44. Rutkowski L, Jaworski M, Duda P. Decision Trees in Data Stream Mining. Studies in big data. 2019 Mar 16;37–50. https://doi.org/10.1007/978-3-030-13962-9_3

45. Tulcanaza Ochoa GF, Cisneros Clavijo PE, Meza Tonato JL, Placencia Guartatanga PG, Manzano Vela MP, Nieto Nuñez FI, et al. Efficacy of wearable cardiac monitoring devices versus traditional methods in detecting atrial fibrillation: a systematic review and meta-analysis. Salud, Ciencia y Tecnología. 2024;4:.962.

46. Saduakas A, Kurakbayev K, Askar Y, Baimuratova M. Duplex ultrasonography for screening and monitoring of carotid artery stenosis for risk stratification of ischemic stroke. Salud, Ciencia y Tecnología. 2024;4:.549.

47. Patra SS, Jena OP, Kumar G, Sreyashi Pramanik, Misra C, Singh KN. Random Forest Algorithm in Imbalance Genomics Classification. In Data Analytics in Bioinformatics: A Machine Learning Perspective. 2021 Jan 18;173–90. https://doi.org/10.1002/9781119785620.ch7

48. Gu X, Angelov PP. Multi-Class Fuzzily Weighted Adaptive Boosting-based Self-Organising Fuzzy Inference Ensemble Systems for Classification. IEEE Transactions on Fuzzy Systems. 2021;30(9):1–1. https://doi.org/10.1109/tfuzz.2021.3126116

49. D Sudharson, S Ashfia Fathima, Kailas PS, K S Thrisha Vaishnavi, S Darshana, A Bhuvaneshwaran. Performance Evaluation of Improved Adaboost Framework in Randomized Phases Through Stumps. 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA). 2021 Oct 8;1–6. https://doi.org/10.1109/icaeca52838.2021.9675739

50. Milad Niroumand-Jadidi, Bovolo F. Extreme gradient boosting machine learning for total suspended matter (TSM) retrieval from Sentinel-2 imagery. Proceedings of SPIE - The International Society for Optical Engineering. 2022 Oct 28;12263:7–7. https://doi.org/10.1117/12.2638465

51. Li Y, Gou J, Fan Z. Particle swarm optimization-based extreme gradient boosting for concrete strength prediction. 2022 IEEE 6th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC). 2019 Dec 1;982–986:982–6. https://doi.org/10.1109/iaeac47372.2019.8997825

52. Pan B. Application of XGBoost algorithm in hourly PM2.5 concentration prediction. IOP Conference Series: Earth and Environmental Science. 2018 Feb;113(1):012127. https://doi.org/10.1088/1755-1315/113/1/012127

FINANCING

The authors did not receive financing for the development of this research.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHORSHIP CONTRIBUTION

Conceptualization: Achraf BENBA, Ahmed IDRISSI.
Data curation: Zaynab BOUJELB.
Formal analysis: Zaynab BOUJELB, Ahmed IDRISSI.
Research: Zaynab BOUJELB.
Methodology: Achraf BENBA.
Project management: Achraf BENBA, Ahmed IDRISSI, El Mahjoub CHAKIR.
Resources: Zaynab BOUJELB.
Software: Zaynab BOUJELB, Achraf BENBA.
Supervision: El Mahjoub CHAKIR.
Validation: Achraf BENBA.
Display: Achraf BENBA, Ahmed IDRISSI.
Drafting - original draft: Zaynab BOUJELB.
Writing - proofreading and editing: Achraf BENBA, Ahmed IDRISSI, El Mahjoub CHAKIR.